Data Analysis 3 - Assignment 2 - Business Report

Perice Prediction for Arbnb Apartments in Melbourne, Australia

Ali Sial

2/9/2022

Introduction

The aim of this report is build price prediction models, which will help a company that operates **small and mid-size apartments hosting 2-6 guests in Melbourne, Australia**. The predicted prices produced as a result of the analysis explained below will form the base for the company to price their new apartments that are going on market soon. The data used for prediction is the **Airbnb prices in Melbourne,Australia.** obtained from **Inside Airbnb**. The prediction models were built on various predictors, such as the type of property and people it accommodates, locality or what unique features it includes, also incorporating the information about the host and reviews. The prediction algorithms used for this analysis are OLS, Cart, Random forest (with and without tuning) and GBM. I also used LASSO, but as a predictor to extract important variables to be used in the models. After conducting the analysis, the model choose for final prediction based on its performance is **Random Forest with Tuning Parameters**. The entirety of this project, including codes and data, is available on my **GitHub** (please click to open the link).

The Data and Cleaning

The raw data is available on Inside Airbnb, for convenience, It has also been uploaded to my GitHub and can be directly retrieved from there (raw data can be found **here**).The data is of cross-sectional nature containing information related Airbnb listings in Melbourne, Australia which was scrapped between January 08, 2022 and January 09, 2022. Raw data includes 17409 observations (unique listings) and 74 variables. The codes for adjusted raw data explained are available on my **GitHub** (please click to open the link).

I begin cleaning the data by removing all unnecessary variables and also made alterations to the information recorded in the variables to make them useful for analysis. As we predicting the property rental prices so price is our target variable which was recorded in local price and had to be converted to USD. The major task involved in the basic cleaning was to address the amenities. All the amenities were record in the same column as a string and I decide to split theses into individual columns.The new created amenities column amounted to 1468, which were the pooled into meaningful categories. Amenities such as TV or WiFi/internet had a lot of columns that resulted due to their unique features such as speed of internet or size of the TV. Upon creating these pooled binary categories only 79 amenities column remained.

The data was further filter to the parameters set by the company supporting this study. Since we are only interested in apartments or housing units accommodating 2 to 6 people, therefore, I filtered the data based on these two requirements. For property type, I selected four categories that are Condo, Serviced Apartments, Loft and Home/Apartment. Other important variables, such as number of bedrooms or review score ratings, that contained missing values were also imputed and a additional flag variable was added where missing values were more than 5%. The cleaning phase ended with variables having no missing values and **802** remaining observations **119** variables.. Codes for data preparation have been uploaded to my **GitHub** (please click to open the link).

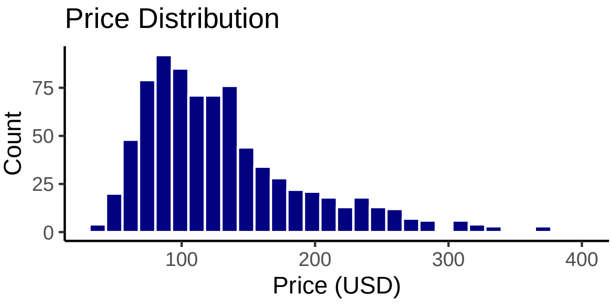
Explanatory Variables

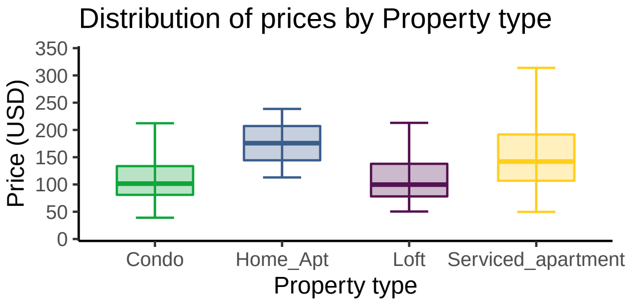
Due to the complexity the variable selection phase was a hulcurian task, eventually based on my understanding I categoriesed these varibles as follows:

- *Factor variables*: Type of property, neighborhood including flag and factorised variable of size variables (such as number of accommodates/bedrooms/beds/baths/minimum nights).
- *Reviews variables*: Review score rating, total number of reviews, total reviews for the property every month, number of days since first review (etc.) including flags.
- *Host variables*: Dummies created for host verification, host being a super host or not, host response and acceptance rates (etc.) including flags.
- *Dummy Amenities*: this included all the binary variables created for amenities being offered at the property.

Exploratory Data Analysis

Upon completion of data cleaning, grouping and feature engineering, I examined the distribution of our target variable i.e. **Price**. The price data, as always, had a right long tail and I felt it was wise to drop few large values. Therefor I filtered the data to observations that had less than USD 400 price. I also decided to use price as is since interpertaion with log normal distribution is complex. Below you can see the graph for price and price based on property type. The interaction terms were also added for the amenities and other binary variables to the models based on their relationship with the property type and impact on prices.





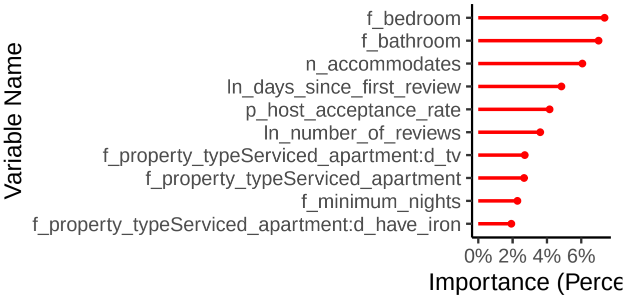
Prediction Modeling

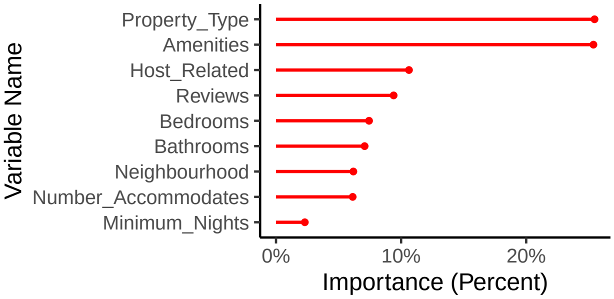
After obtaining the data set that was fit for regression analysis, the data was further split in to train and test smaples. Since I had limited data to work with, to maximise the performance, I divided the observations by adding 70% to train and remaining to test. These samples were used in the machine learning models which are OLS, Random Forest with and without tuning, CART and GBM.

OLS models were built using LASSO as a predictor. Initially I built the most complex model (Model 4) which included the maximum number of explanatory variables and interaction terms. I used this over-fitted model to run LASSO which helped in identifying the key predictors and interaction terms (variables with non zero coefficients).

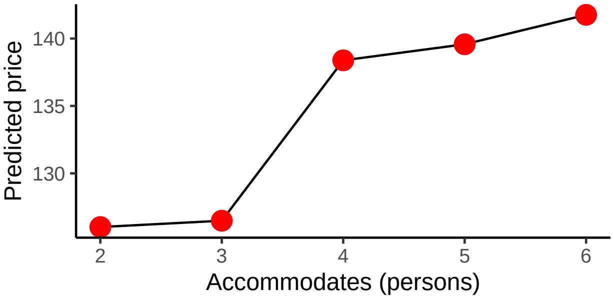
Based on the results produced from 5-fold cross validated Root Mean Squared Error, prediction model random forest with tuning parameters performed the best. Therefore, I decided to select that model further validation and testing. For tuned FR, the maximum number of trees was set at 500, the lowest RMSE value was 48.05 which was obtained with 5 terminal nodes and 12 variables in each node.

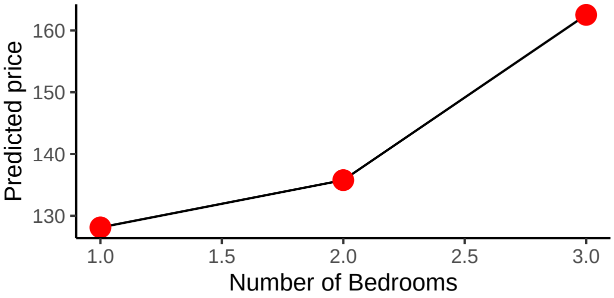
Variable Importance: The purpose of variable importance is to identify the predictors that impact the target variable the most. In this study we used the best performing model which is Random Forest with tuning to identify these variables. Number of bedrooms, number of people the property accommodates and number of bathrooms were the top performing variables. Below you can also see the graph highlighting the top 10 variables. Moreover, the second graph show the grouped explanatory variables importance graph which clearly shows that overall amenities and property type tend to be the most impact on the price.





Partial Dependencies and Sub Sample: Based on the results from the variable importance plot we analyse the shape of association between average y and important x variables, condition on the rest (BEKES, 2022). To do this I decided to take two most important variables: number of people accommodated by the property and number of bedrooms. Below you can see the partial dependency plots for these variables. For both variables the PDP rather shows a fairly linear relationship with predicted prices. Using the Sub Sample approach, I predicted prices using important x variables. Based on the result of sub sample, we can say the company should invest in serviced apartments because it had the lowest error and highest predicted price i.e. $149.32.





Conclusion

Analysis performed for predicting price results in Random Forest model with tuning performed the best since it produces the lowest RMSE value. However, due to the limitation of Random Forest because it is considered as a black-box model, therefore, it is usually difficult to explain the regressions to anyone not familiar with it. Even though, to avoid complexity, I recommend to rather select an OLS model, which in this case was model 3. The RMSE difference between Random Forest Auto Tune and OLS (model 2 since that was selected for comparison) was approximately $10 which is alot to be disregarded. Thus, I will stick with selecting random forest as the preferred model. Details other algorithm (CART and GBM), please refer to the technical report.